

Text-based Hierarchical Image Classification and Retrieval of Stock Photography

Anna Bjarnestam

Getty Images Ltd
101, Bayham Street
London, NW1 0GA, UK
anna.bjarnestam@getty-images.com

Abstract

Tony Stone Images have designed a textual classification structure and an image retrieval system to store and retrieve pictures within the stock photography domain. The image retrieval system has been designed from the stock photography domain-specifics with facilities to search for different objects and their attributes within an image, and on relations between these objects. The retrieval system is text-based with a set of 12, 000 controlled hierarchical keywords in a semantic network structure which has approximately 45,000 additional search terms in the form of synonyms linked to these keywords. The keywords allow for searching both on specific objects within a picture, and for abstract concepts that an image signals as a whole. The keywords are added manually to each image and they are searched on in different fields of the user interface of the retrieval system depending on what type of content is wanted. Furthermore, the image retrieval system deals with ambiguity by prompting the user at the time of searching and it allows for multi-lingual searching in French, German and American English besides British English.

1 Introduction

Getty Communications is an international provider of visual content and the group currently comprises Tony Stone Images (contemporary stock photography), Photodisc (contemporary stock photography), Allsports (sports imagery), Hulton Getty (archival imagery), Liaison Agency (news and photojournalism), and Energy (contemporary stock footage). The Tony Stone Images collection comprises more than 1 million contemporary stock photographs which are licensed for limited use across sectors as diverse as Advertising, Design, Corporate Communications, Editorial, Sales Promotion, Publishing, Multimedia, Direct Marketing, and On-line Services. Tony Stone Images is currently in the process of digitizing their imagery and to date there are about 70, 000 scanned pictures.

1.1 Types of Image Content and the Stock Photography Domain-specifics

What an image *means* or what its *message* is vary depending on what the purpose with the image is [14]. For example, within news and archival imagery, then matters like *time*, *place* and *people involved* are important rather than what *objects* are visible in the pictures [11, 13]. On the other hand, contemporary stock photographs have a completely different purpose, that is, they are often used in advertising and should be applicable to many types of products, services and so on. Often, one and the same picture can be used for two different markets, for example, the cigarette industry and the travel industry, since the image does not necessarily refer to particularities but, rather, it can signal an *abstract message* of, for example, *peace* and *relaxation*. Of course, what we are dealing with here are not abstract linguistic expressions but *image content* which unfortunately is a vague term and researchers within both text-based and content-based image retrieval use a variety of terms when they talk about the *content of images* [9, 14]. Tony Stone Images have done small scale in-house surveys on what image content clients search for, that is, analyses of linguistic expressions clients have used when asking for pictures and the findings show that users often request image content by using abstract terms, for similar surveys see Enser [12, 13] who analysed user queries from Hulton Getty, and Keister [15] who describes the query analysis forming the basis for a still picture retrieval system at the National Library of Medicine. What is important is that these abstract terms cannot be inferred from the visible parts or objects in an image. Thus, these abstract concepts cannot be added automatically by defining rules of what objects, or parts, an image contains. This problem has also been acknowledged by researchers in the field of automatic indexing and retrieval [9, 10].

There are, of course, different types of information to be found at different levels in a picture besides the abstract information [13, 19]. *Primary* information can be matters like forms, shapes, light intensity, texture and colours and the relative positions of such phenomena, all of which are criteria researched within the field of content-based image retrieval [5]. However, in contemporary stock photography users talk about pictures using words rather

than shapes, forms, intensity, texture and colours. A client may ask for a picture of a *happy person of a certain age and gender, who is sitting by a table holding a cup of coffee in one hand and a newspaper in the other*, and the whole image should signal *relaxation*. Such interpretations, that is, both of what objects and attributes are visible and of what the image signals as a whole, require the use of words - abstract and concrete - in the classification and retrieval processes (Cf [13] on the relationship between the visual and linguistic modes in information retrieval with the conclusion that the linguistic-linguistic mode is not effective for image retrieval). Clients do not always know exactly what objects they would like to retrieve, but they often know what an image should convey as a whole, and, thus, it is difficult to see a situation where an image storage and retrieval system within the stock photography domain makes use of the 'similar picture' or the 'sketch pad' based on feature identification [1, 2] without any linguistic mode in classification or retrieval. However, to have both visual retrieval and text-based search facilities with the possibility to search on abstractions seem like an ideal mix [4], although difficult to implement.

There are also knowledge-based image understanding systems [3] which automatically analyse what basic objects appear in an image by mapping numbers into symbols in various ways, for example, by performing quantification analysis [6] or through rule-based recognition-independent segmentation [7, 8] which works well in well-defined domains. However, due to the fact that the database of images underlying the Tony Stone image storage and retrieval system is not constrained to a particular knowledge-domain, it was believed that developing any such automatic object recognition would be too costly to be validated. Also, since many of the photographic techniques used for some of Tony Stone's images, such as blurring, soft focus, distortions, various close-ups and so on, there are objects that look very similar although they are in fact completely different, which is shown in Figure 1. The left picture is a light micrograph of a buttercup root showing central vascular bundles, and the right picture is a globe made up of an aerial view shot of crowds of people and still, both pictures show similarities in colour, texture and shape. By manual keywording of such pictures, the correct image content is easily recognised. Currently, the manual application of keywords to one image takes approximately 7 minutes which was considered a cheaper, more efficient and reliable method than to create an object-recognition rule base for automatic application.



Figure 1: Two different objects with similar features

1.2 Former Handling Process and Goals of the Image Retrieval System

Most of the processes involved in the indexing, classification, and retrieval of Tony Stone images are based on a manual tradition. First, a seventy character text caption is added to every image which describes its content in full text as a sentence. To this, an addition of 350 characters can be added as a blurb if there is a need to add scientific information. A simple solution to the retrieval needs could, thus, be to add a simple free-text search engine for search on the captions. Still, there are a number of difficulties with free-text retrieval systems, and most of them are due to the difficulties of creating a system that can perform linguistic analysis in a meaningful way which Warner discusses in her paper on the relationship between linguistics and information retrieval [20]. One of the main problems with a free-text retrieval system is ambiguity. By ambiguity is meant that in a free-text search engine, a user searching on, for example, the term *orange* will retrieve both pictures of the fruit *orange* as well as pictures that have the colour *orange*. This is frustrating for the user who may only want pictures of the fruit *orange* and it may be time-consuming to go through all the orange images and select only the ones of fruits, if searching is done in a large collection of images. One of the goals was, thus, to create an unambiguous retrieval tool. The kind of ambiguity

referred to here is *synchronic*, as opposed to the type of *diachronic* ambiguity Busch [16] mentions in his paper on historical image classification, where ambiguity arises from the fact that objects, events and situations change over time and can thus have several different meanings at any point in time. Another goal was that the image retrieval system should be multi-lingual and cater for French, German and American English besides British English and ambiguity should be dealt with in each language.

The project team tested many image retrieval, indexing and classification systems, both commercial and academic, but no system was found to be perfect for the specific needs of stock photography. Most available commercial tools were designed for document-imaging (DIP) [9] of which some make use of Natural Language Processing tools to automatically analyse, search and retrieve natural language [20]. Some of the available methods and tools were found to be useful if they could be implemented with the core classification needs of the stock photography domain.

Another important task of the image retrieval system is to speed up the process of giving clients information about what pictures there are in the collection that meet their requirements. The traditional process of operating consists of a client calling a sales person with a query which is passed on to a picture researcher who by memory or by looking in particular topic boxes with transparencies arranges a selection of pictures which are then sent out to the client. If the client wants to make a purchase, s/he has to call back to the sales staff. The image retrieval system is designed to help sales staff so that they are able to retrieve a selection that is based on the *whole* scanned catalogue and which better meets the client's requirements, quicker than before. In other words, effectiveness, relevance and recall of selected pictures are improved, and time is saved. Since the collection is constantly growing, there was a need to move away from reliance on memory. An additional goal of the image retrieval system is to give clients direct access to pictures to a greater extent than what was possible before. The image retrieval system extends this access by being the core for CD-ROMs, and from September 1998 also for the Internet.

1.3 Image Retrieval System Structure

The image storage and retrieval system was created by scanning each image in the database at low resolution and storing them as JPEGs. Compressed thumbnails were then created to be retrieved at the time of searching to speed up the searching process so that the actual image file is only accessed when the user clicks on a single image of the thumbnail result list. The keywords are stored alongside the image files as comma separated text files, corresponding to the linguistic-linguistic model of visual information retrieval as described by Armitage and Enser [22] and Enser [13]. The keywords are stored in different database fields to allow the separation of different objects, their attributes and the relations between them, rather than storing them as an unstructured list of words linked to each picture. Below, the input keyword process, the keyword structure, and the retrieval process are described.

1.3.1 Keyword Structures and Database Fields

The keywords, currently 12, 000, are divided into nine semantic hierarchical tree structures, namely the: *Geography*, *Natural* (consisting of *Animals*, *Natural Phenomena*, and *Flora*), *Man Made*, *Human Emotions*, *Image*, *Concepts & Topics*, *People*, *Activity* and *Description* trees. Each word has its specific position in the network structure with the name of the individual trees as the top nodes. To take an example of a string in the keyword hierarchy, the word *Dog* has the string: *natural/living organism/animal/mammal/pawed mammal/dog family/dog*, that is, it lives in the natural tree and has the term *Dog Family* as its immediate parent. *Dog* has as a sibling *Wild Dog* and as daughters *Pure Bred Dog*, *Mongrel* and *Stray Dog*. The keyword hierarchies were constructed as an attempt to achieve maximum usefulness of minimum input. In other words, since the keywords live in semantic networks, there is no need to say that a *Pure Bred Dog* is a *mammal* and an *animal*, since that information is inferred from the parent terms in the hierarchy where the word lives. No existing iconic classification scheme was used, such as ICONCLASS or AAT [17, 18], because these are mainly aimed for the Art and History domains, rather than the contemporary stock photography domain.

The procedure of adding keywords to an image begins by the addition of keywords for the individual objects in a picture, that is, the primary subject matter [13]. The words used for object description can be taken from all nine tree structures. Then a set of *attributes* relating to each object are added and these words are linked to the object they describe. Attributes of an object are not only what in visual information retrieval are called *visual features* [5] such as colour, texture, shape, and so on, but also other features that humans typically associate with a particular object such as age of people, hair colour, eye colour, clothing, size, what an object is used for and so on. To be able to search on relations between these objects a link facility was set up. To take an example, an image of *a young girl holding a toy bucket* may have the object of a *person* with the attributes *female*, *2-3 years*, *blond*, *blue-eyed*, *beige coat*, and another object of a *toy* with the attributes *bucket* and *yellow*. A link can be created by the word *holding* so that Object 1, the girl, is *holding* Object 2, the toy. This feature allows a search on "*Girl Holding Toy Bucket*" to result in only the pictures where a girl is actually holding a bucket, and not pictures where, for example, the bucket is

on the ground or someone else is holding it. The link words can also be taken from any of the nine trees but are usually taken from the Activity tree which contains most words describing spatial positions, states and activities, that is, verbs and prepositions.

Besides the objects and their links, information about the setting (or background) is added which can be geographical information, if known, or locations such as *beach*, *outdoors*, *kitchen* and so on. Basically, the information added in *the Setting field* denotes anything that is not prominent in the picture. If a location is considered prominent in the picture, then that location is made an object of its own rather than being added in the Setting field. Other types of information added in the setting field are, for example, objects that can be seen in the background of an image, that is, those objects which are not the main objects of interest. Furthermore, photographic terminology is added in a separate database field called *the Image field* to define matters such as, for example, *vertical/horizontal image*, *black & white* or *colour image*, *image technique* used (for example *soft focus*, *blurred*, *scientific imaging*), *image manipulation* (for example *hand tinted*, *composite*) and so on.

The *Concepts & Topic* words are abstract in meaning and/or are words that refer to the whole picture rather than to specific objects and attributes within a picture. Such words are added in yet another separate field in the image database. There is at least one concept or topic added to every picture but usually there are about 5-10 on each picture. Topics are words for various knowledge-domains such as *Lifestyles*, *Science*, *World Travel*, *Health & Beauty*, *Nature*, *Business*, *Education*, *Ideas* and so on. Concepts, on the other hand, are abstract words that can be added to pictures from all types of knowledge-domains such as *Happiness*, *Strength*, *Thought*, *Time*, *Newness*, *Opposites* and so on, and they are words corresponding to the 'iconography and iconology modes' of Panofsky's image analysis levels [19].

New words can be added in the keywording input process and can later be rejected or converted into new keywords, which means that the keyword trees constantly change. Furthermore, to each keyword there is a set of synonyms added which provide an additional route to the controlled vocabulary. These synonyms have been selected specifically for the imagery to make sure that they all lead to relevant images. All plural forms of nouns are stored as synonyms, as well as words that are close in meaning to the keywords. For example, to the abstract concept *Happiness* the synonyms *Merriment* and *Delight* are added among other terms, while to the concrete term *Happy*, which is a term that can be added to an object within a picture, the synonyms *Looking Happy* and *Looks Happy* are added among other terms. You can talk about someone or something looking happy, that is, this can be inferred from visible criteria, while the more abstract terms *Happiness*, *Merriment* and *Delight* are used to denote *situations* rather than to denote individual objects and an image as a whole is a kind of situation.

All keywords have been translated into French, German and American English and synonyms in these languages have also been added. This multi-lingual component is similar to that created at Nanyang Technological University [4]. Small scale in-house surveys showed that although there are linguistic differences among keywords in different languages, the general abstract interpretation of pictures as a whole seems to be the same among native speakers of the four translation languages, which is a contradiction to what many researchers in image retrieval claim [5, 10, 13].

Furthermore, there are created terms called *Roof Terms* and *Macros* which are stored search formulae consisting of existing keywords and Boolean operators. These Roof Terms and Macros are not part of the hierarchical structure but are made up of keywords from different parts of the trees and do, thus, not follow the hierarchical inclusion of meaning. It is, thus, possible to keep a hierarchy of controlled terms and still have a set of non-hierarchical terms taken from different trees. Examples of Roof Terms are *Endangered Animals*, *European Capitals*, various currencies and other terms that are often requested by clients. Macros are terms which can be split into objects and attributes such as *Kitten*, *Lamb* (which are objects of *Cat* and *Sheep* respectively, with an attribute of *Young*) or herds of various animals and so on. Synonyms are added also to Roof Terms and Macros.

1.3.2 Searching Process and Query Language

The query language is based on the keywords and synonyms, and the Boolean operators AND, OR and NOT. A search containing a synonym is automatically translated into the keyword it links to, before the search query is processed. It is possible to search in different fields corresponding to the fields in the underlying database and there are three different screens depending on the complexity of the query (a simple, an advanced and a maintenance screen). On all screens there is one field that searches across all fields and there are fields for meta-data searches such as photographer, image reference number, user restrictions and so on. There is also a full vocabulary browser on all screens with a drag-and-drop facility (see Figure 2). On the advanced screen it is possible to search for specific objects and attributes and on relations between these objects. In Figure 2 a screen shot of an advanced search is shown. In the Object 1 box a *Person* Object is defined with the attributes of *Female*, *Child*, *Solo* and *Playing*. In the Object 2 box the object of a *Doll* is defined and in the Object 3 box a *Swing* is added. Below these

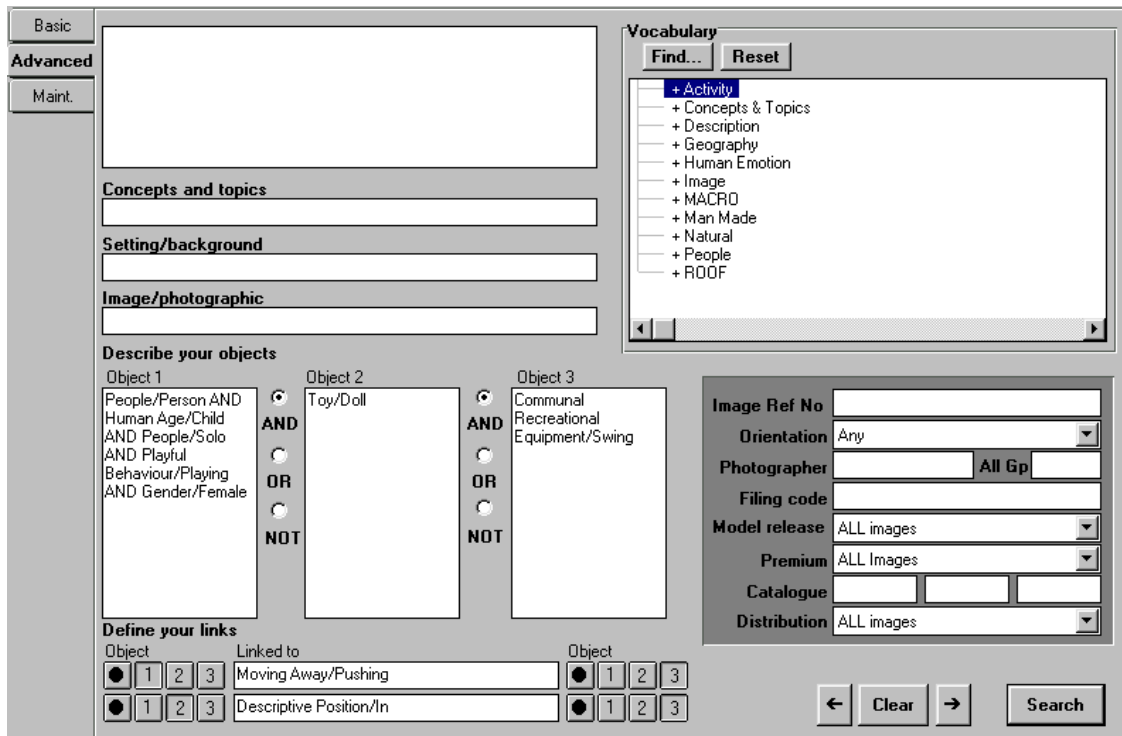


Figure 2: An Example of Searching on Two Linked Objects

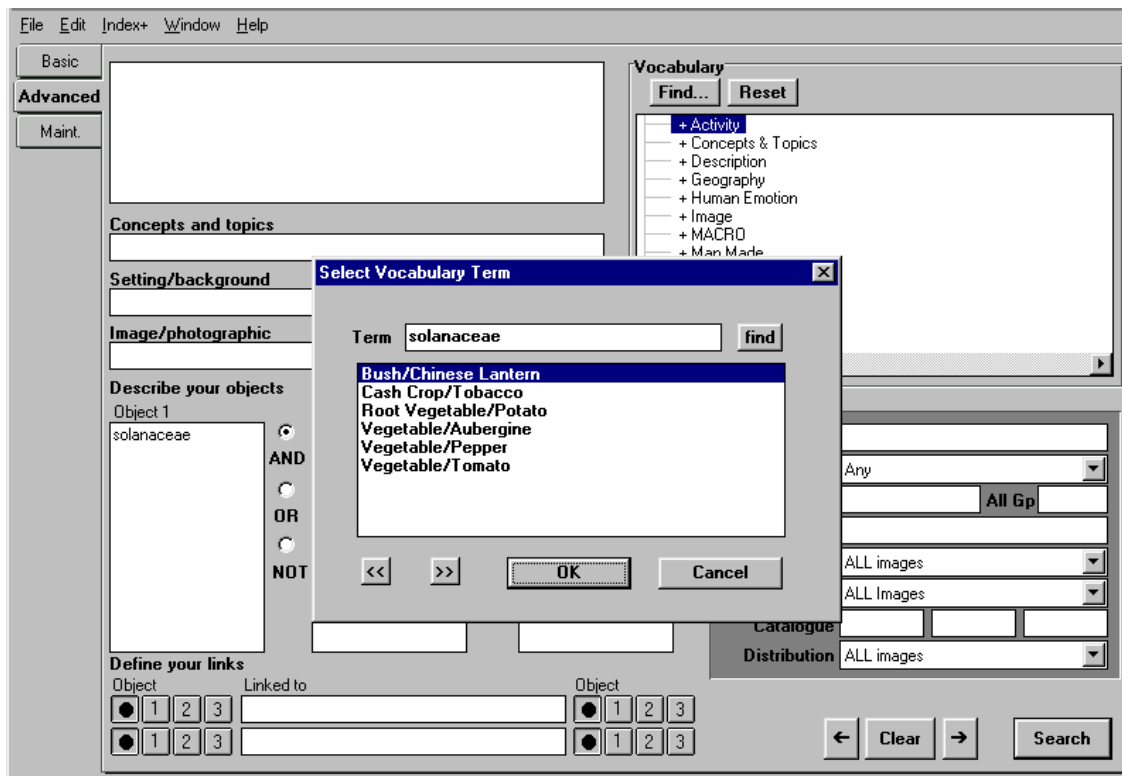


Figure 3: Interactive User Prompt to Disambiguate Word Meanings

boxes the top Link box has *Object 1* linked with the term *Pushing to Object 3*, and in the link box underneath, *Object 2* is linked with *In to Object 3*. This search will result in the retrieval of all pictures of girls who are alone in the picture and who are pushing swings with a doll inside. Thus, no irrelevant pictures will be retrieved such as a picture of a group of children where one child is in a swing and another one is holding a doll.

Ambiguity of words is dealt with in all languages by simply defining the immediate parent to an ambiguous word in the hierarchy and prompting the user with the set of choices an ambiguous word gives. Figure 3 shows what a prompt may look like. In Figure 3 a search on the Latin family name *Solanaceae* caused the disambiguator to prompt the user with a set of alternative keywords which are all part of that Latin family, namely *Chinese Lantern, Tobacco, Potato, Aubergine, Pepper* or *Tomato*, and the user can select one, two or more, or all of these terms. This prompting takes place before the actual search is done.

The result screens are thumbnails with or without the caption texts and from there a user can go to a single image view and zoom in and out of the image, and view all keywords and detailed data attached to that image. A user can also view the hits as a result list of caption texts, reference numbers and other criteria, as well as saving the searches either to a file or by selecting images to a lightbox. It is also possible to print the images, lists or lightboxes in various formats.

1.3.3 Evaluation of the Image Retrieval System

Although it is very difficult to compare different image retrieval systems since there are no set criteria for degree of relevance of search results, recall or indeed what constitutes a useful image retrieval system [5], an attempt is made to describe the usefulness of the image retrieval system in question and the following points will be discussed: *Recall, Relevance* and *Ease of Use*.

To begin with recall, the image retrieval system brings back a set of pictures matching the user's query from the whole catalogue of stored images. Since the keywording of images was done manually, a search on any keyword will bring back every image in the catalogue that matches that image content. The keyworders have deliberately added a keyword as an *Object* if the matter in question was considered to be prominent in the picture or s/he has added it in the *Setting* field if it was considered to be less important in the picture. Cawkell [9] claims that recall cannot be ideal for text-based keyword systems since humans have different opinions on what objects in a picture are important and on what abstractions are conveyed (Cf Enser [13] and Shatford [21]). Enser [13] even say that one and the same person may interpret an image differently at different times. To try to overcome this problem, the Tony Stone keywording is done by adding as many visible objects and abstract keywords in a picture as possible, which means that a keyworder has to think about an image from many different angles. Quality control is done to make sure that all necessary image content is keyworded, and recall have been found to be sufficient for the purposes of retrieving image requests from clients.

Secondly, images retrieved are relevant for two reasons: partly because of the classification structure of image content into objects and attributes (Object fields), photographic terminology (Image field), abstract terminology (Concept & Topic field) and background information (Setting field), and partly because of the fact that the system prompts the user when a requested term is ambiguous. Thus, in cases where the system does not 'know' what content a user is looking for, the system asks the user to select which term(s) s/he wants. Approximately 1/5 of all keywords and synonyms are ambiguous and without these prompts the user would retrieve many irrelevant and incorrect matches. If a user knows exactly what objects s/he wants to retrieve, it is easy to define a search and retrieve only the relevant pictures. If, on the other hand, a user would like to see suggestions of images where the actual objects are less important, s/he can search on the abstract Concepts & Topics and browse the results.

Thirdly, the retrieval system is fairly easy to use since the relationship between the classification approach and the retrieval method matches the prototypical user query. An in-house investigation of 1000 user queries was done which found that the prototypical stock query consists of a noun phrase incorporating a noun with a nonfinite postmodified clause. To describe such a query without going into linguistic analysis details, a few example queries are given: *Ant carrying leaf; Baby walking in park; Man drinking a cup of coffee*. Although some users do not like to use Boolean operators, all these queries can easily be constructed by simply adding an AND between each content word (that is, words carrying meaning), for example: *Ant AND Carrying AND Leaf*. If the user has some experience of boolean operators and of the various search fields of the advanced screen, s/he can create a query by placing the first noun (that is 'Ant', 'Baby', and 'Man' respectively) of the exemplified queries in an Object box, and then place the second noun of each query (that is 'Leaf', 'Park', and 'Cup' with the attribute 'Coffee' respectively) in a second Object box. Then, the user can add the non-finite verb (that is 'Carrying', 'Walking', 'Drinking' respectively) in a Link box to link the two objects together so that only relevant images are retrieved. For the query 'Ant Carrying Leaf', the user only wants images where an ant is actually carrying a leaf, and not images where an ant is carrying something else and a leaf is just visible somewhere in the image. Although the classification structure was not set up with the specific needs of an image retrieval system in mind, it is fairly easy to map a user's request to a search query

and retrieve relevant results. Evaluating image retrieval systems is not a simple task [5] but it is a task that has to be done by comparing how the system performs in relation to the purpose of the system which will, of course, vary depending on the image domain in question. For the stock photography domain, where a great many clients come from the Advertising domain and where there is a need to classify and search on both concrete objects and abstract image content such as what an image as a whole signals, the image retrieval system discussed in this paper performs well.

2 Conclusion

Tony Stone Images has developed a text-based image retrieval system that meets the specific needs of users within the stock photography domain. The main features of the image retrieval system are the underlying textual hierarchical classification structure, the separate database fields for concrete and abstract textual image content, disambiguation through prompts to the user, and the fact that the retrieval system is multi-lingual, which allows searching in French, German and American English besides British English. Currently, investigations as to what may be further added to the image retrieval system are being made, and in particular whether it is possible to add automatic components such as a 'sketch pad' and automatic feature identification and/or Natural Language Processing tools to give users an even greater choice of search possibilities. The image retrieval system will also be the core of a client access retrieval system on the Internet for viewing, selecting and purchasing pictures online to be launched in September 1998.

3 References

1. Del Bimbo, Alberto; Pala, Pietro. Visual Image Retrieval by Elastic Matching of User Sketches, In: Institute of Electrical and Electronics Engineers (IEEE) Transactions on Pattern Analysis and Machine Intelligence, Vol 19, No 2. IEEE Computer Society. 1997 (pp121-132)
2. Ramesh, Nagarajan; Sethi, Ishwar. Feature Identification as an Aid to Content-based Image Retrieval, In: Storage and Retrieval for Image and Video Databases III, SPIE Proceedings, San Jose, 1995, Vol 2420: 2-11
3. Crevier, Daniel; Lepage, Richard. Knowledge-based Image Understanding Systems: A Survey, In: Computer Vision and Image Understanding, Vol 67, No 2, August 1997: 161-185
4. Jay Kuo, C.-C. Multilingual Information Retrieval System, In: Multimedia Storage and Archiving Systems, SPIE Proceedings, Boston, 1996, Vol 2916: 33-44
5. Gupta, Amarnath; Jain, Ramesh. 1997. Visual Information Retrieval, In: Communications of the ACM, 1997, Vol 40, No 5: 71-79
6. Meisels, A. Levels of Knowledge for Object Extraction, In: Machine Vision and Applications, 4, 1991: 183-192
7. Kanade, T. Survey-region Segmentation: Signal Versus Semantics, In: Computer Graphics and Image Processing, 13, 1980: 279-297
8. Nazif, A.M; Levine, M.D. Low Level Image Segmentation: an Expert System, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, Vol 6, No 5: 555-577
9. Cawkell, A.E. Yardsticks for Picture Databases. British Library R&D Report 6116, The British Library Board, 1993
10. Cawkell, A.E. Indexing Collections of Electronic Images: A review. British Library Research Review No 15, 1993
11. Enser, P.G.B.; Orr, H.P. Creative Picture Research Using the Hulton Deutsch CD Collection. Program, Vol 27, no 4, October 1993: 389-398
12. Enser, P.G.B. Query Analysis in a Visual Information Retrieval Context. Journal of Document and Text Management, 1993, 1(1): 25-52
13. Enser, P.G.B. Pictorial Information Retrieval. Journal of Documentation, 1995, 51(2): 126-170

14. Evans, Hilary. *Practical Picture Research: A Guide to Current Practice, Procedure, Techniques and Resources*. Blueprint, London, 1996
15. Keister, L.H. *User Types and Queries: Impact on Access Systems*. In: Fidel, Raya et al (Eds). *Challenges in Indexing Electronic Text and Images*. ASIS Monograph Series, Medford, New Jersey, 1994: 7-22
16. Busch, J.A. *Thinking Ambiguously: Organizing Source Materials for Historical Research*. In: Fidel, Raya et al (Eds). *Challenges in Indexing Electronic Text and Images*. ASIS Monograph Series, Medford, New Jersey, 1994: 23-56
17. Van de Waal, H. *Iconclass – An Iconographic Classification System*. North Holland Publishing Company, Amsterdam/Oxford/New York, 1985
18. Petersen, T. *Developing a New Thesaurus for Art and Architecture*. *Library Trends* 1990, 38(4): 644-658
19. Panofsky, E. *Meaning in the Visual Arts*. Penguin, London, 1993
20. Warner, A. *The Role of Linguistic Analysis in Full-Text Retrieval*. In: Fidel, Raya et al (Eds). *Challenges in Indexing Electronic Text and Images*. ASIS Monograph Series, Medford, New Jersey, 1994: 265-276
21. Shatford, S. *Analyzing the Subject of a Picture: a Theoretical Approach*. *Cataloguing & Classification Quarterly*, 5(3), 1986: 39-61
22. Armitage, L.H; Enser, P.G.B. *Information Need in the Visual Document Domain*. Report on Project RDD/G/235 to the British Library Research and Innovation Centre, August 1996